RADIOLOGICAL EVALUATION OF THE INTERFACES AFTER CEMENTED TOTAL HIP REPLACEMENT
INTEROBSERVER AND INTRAOBSERVER AGREEMENT

A. W. McCaskie, A. R. Brown, J. R. Thompson, P. J. Gregg

From Glenfield Hospital, Leicester, England

Three radiological methods are commonly used to assess the outcome of total hip replacement (THR). They aim to record the appearance of lucent areas and migration of the prosthesis in a reproducible manner. Two of them were designed to monitor the implant through time and one to grade the quality of cementing. We have measured the level of inter- and intraobserver agreement in all three systems.

We randomised 30 patients to receive either finger packing or retrograde gun cementing during Charnley hip replacements. The postoperative departmental radiographs were evaluated in a blinded study by two orthopaedic trainees, two consultants and two experts in THR. The trainees and consultants repeated the exercise at least two weeks later. We used the unweighted kappa statistic to establish the levels of agreement.

In general, intraobserver agreement was moderate but interobserver agreement was poor, with levels similar to or less than those expected by chance. Our results indicate that such systems cannot provide reliable data from centres in different parts of the world, with various levels of surgeon evaluating radiographs at differing time intervals. We discuss the problem and suggest some methods of improvement.

Received 19 June 1995; Accepted after revision 26 October 1995

Outcome measurement has become an increasingly important part of orthopaedic research and the methods used must be reliable (Gartland 1988; Wright and Feinstein 1992). In hip replacement one method of assessing outcome is to determine radiological signs of loosening which include migration, defects in the cement mantle and radiolucent zones at implant interfaces.

Attempts have been made to standardise the assessment of radiographs so that rates of radiological loosening at different centres can be compared. Gruen, McNeice and Amstutz (1979) developed the widely-used system in which the femoral component interface is considered in seven zones (Fig. 1); these allow the location of cement fractures and of lucent lines either at the cement-bone or the cement-prosthesis interface. The Gruen zones are also the basis of the more comprehensive system of Johnston et al (1990) which incorporates both clinical and radiological

Fig. 1

Diagram to illustrate the zone system used by Gruen et al (1979) and Johnston et al (1990).
observations including heterotopic ossification and stem position.

Barrack, Mulroy and Harris (1992) emphasised the ‘quality’ of cementing in hip replacement, and described four grades in postoperative radiographs: A, complete filling of the medullary cavity by cement, a so-called ‘white-out’ at the cement-bone interface; B, slight radiolucency at the cement-bone interface; C, radiolucency involving 50% to 99% of the cement-bone interface or a defective or incomplete cement mantle; and D, definite radiolucency at the cement-bone interface of 100% in any projection, or a failure to fill the canal with cement such that the tip of the stem is not covered.

These three systems of assessment have been used as outcome measures to demonstrate the effectiveness of a given intervention, such as design of prosthesis or technique of fixation. We have been unable to find any studies of the levels of inter- and intraobserver agreement in these assessments.

PATIENTS AND METHODS

We performed a randomised, prospective study of 31 patients undergoing primary Charnley cemented arthroplasty. All patients gave their informed consent and were allocated randomly to have the femoral cement introduced either by the finger-packing technique (15 prostheses) using normal viscosity cement (CMW Laboratories, Exeter, UK) or by a retrograde gun technique (16 prostheses) using reduced viscosity cement (CMW Laboratories, Exeter, UK). All patients had brushing and lavage of the canal and the use of a Hardinge cement restrictor (DePuy International, Leeds, UK). In all other respects the technique was consistent. Postoperative radiographs were taken during the week after surgery.

One set of radiographs was not available for assessment, leaving 14 hips in the finger-packing group and 16 in the retrograde group. We designed a proforma which included the three outcome measures and allowed an observer to record findings. The anteroposterior radiograph of each hip was assessed blindly by six orthopaedic surgeons, two trainees, two consultants and two experts with special knowledge of hip replacement. The radiographs were coded and identifying features were masked. None of the reviewers was responsible for the care of the patients. The trainees and the consultants were asked to review the series on two separate occasions at least two weeks apart. Review was performed on an individual basis with the same instruments and with no external advice, and on the second viewing the code and the order were changed. The examiner was allowed to reject any film considered not to be of sufficient quality.

Inter- and intraobserver agreement was assessed by the unweighted kappa statistic (Cohen 1960) which measures the extent of agreement over that which would be expected by chance. On an arbitrary but widely-used scale suggested by Landis and Koch (1977), six levels of agreement are recognised depending on the value of kappa: poor (< 0.0), slight (0.0 to 0.20), fair (0.21 to 0.40), moderate (0.41 to 0.60), substantial (0.61 to 0.80) and almost perfect (0.81 to 1.00). Calculations were performed using a procedure written by Rook (1994) for the GENSTAT statistical package. This calculates a generalised kappa statistic (Fleiss 1971), which is suitable for situations in which there are more than two observers.

To measure intraobserver variability we calculated the kappa statistic for each of the four observers who studied the radiographs twice and summarised them as a single average.

To measure interobserver variability we calculated the generalised kappa statistic for the first readings of all six observers and then separately calculated it for the second readings of the four who reassessed the films. These two figures were summarised as a simple average.

To compare the grade of observer, namely trainee, consultant, and expert, we calculated the kappa statistic for the two people at each grade. When the pair of observers had assessed the radiographs on two occasions (trainee and consultant), the value of kappa was calculated for each occasion and summarised as a simple average.

RESULTS

Gruen system. Both the cement-bone and prosthesis-cement interfaces were studied. To simplify the assessments under the Gruen system observers were said to agree if they both identified a gap, even if they disagreed about its location. This reduced each observer’s responses to a three-point scale of gap, no gap and unreadable.

For the stem-cement interface the intraobserver agreement was moderate with individual kappa statistics of 0.69, 0.38, 0.52 and 0.71, giving an average agreement of 0.57 (95% confidence interval (CI) 0.32 to 0.82). Interobserver agreement was only fair: the observers agreed on 72% of occasions, a little more than would be expected by chance, so that the average kappa was 0.28 (95% CI 0.12 to 0.44).

For the cement-bone interface the results were worse. The intraobserver kappa statistics were 0.31, 0.55, 0.23 and 1.00 giving an average of 0.52 (95% CI 0.12 to 0.92). The interobserver agreement was very slight with agreement on only 53% of occasions, little more than would be expected by chance. The average kappa was 0.02 (95% CI –0.10 to 0.14). The interobserver agreement for each grade of reviewer was calculated: for trainees agreement was fair, the average kappa was 0.373 (95% CI 0.06 to 0.69). For the consultants it was slight, the average kappa was 0.135 (95% CI –0.18 to 0.48). For the experts agreement was also slight; the average kappa was 0.133 (95% CI 0.32 to 0.55).

Barrack system. The results are shown in Table I. This method of assessment allows selection from four grades (A
to D), but in our study grade D was never used. Since observers were allowed to reject a radiograph as unreadable they could therefore give one of four answers: A, B, C or rejected (R). The four readers who assessed the radiographs on two occasions showed variable levels of intraobserver agreement; kappa 0.07, 0.55, 0.43 and 0.63, respectively. The average intraobserver agreement was moderate, with kappa 0.42 (95% CI 0.18 to 0.66). Interobserver agreement was poor. Observers agreed with one another on only 30% of occasions which is less than would be expected by chance. The average kappa was –0.04 (95% CI –0.10 to +0.03).

The interobserver agreement for each grade of reviewer was also calculated. For the trainee agreement was slight, the average kappa 0.052 (95% CI –0.15 to 0.26). For the consultant grade agreement was poor with some obvious disagreement, the average kappa being –0.483 (95% CI –0.70 to –0.23). Finally, for the expert grade agreement was slight; the average kappa was 0.18 (95% CI –0.10 to 0.47).

**Johnston system.** This system of assessment is the most complex of the three. We therefore studied only lucency at the cement-bone interface. As with the Gruen scale, agreement was said to have occurred whenever two observers agreed that there was a gap, even if they differed in their assessment of its location or size. Each assessment was reduced to a three-point scale of gap, no gap and unreadable. Intraobserver agreement was variable with kappa statistics of 0.63, 0.51, 0.44 and –0.09, respectively, giving a fair average of 0.37 (95% CI 0.04 to 0.70). Interobserver agreement was poor. Observers agreed on 46% of occasions, less often than would be expected by chance alone. The average kappa was –0.08 (95% CI –0.18 to +0.03).

The interobserver agreement for each grade of reviewer was poor for the trainee grade with an average kappa of –0.17 (95% CI –0.49 to +0.15). For the consultant grade, agreement was also poor; the average kappa was –0.52 (95% CI –0.78 to –0.26). For the expert grade, agreement was slight with an average kappa of 0.06 (95% CI –0.26 to +0.37).

**DISCUSSION**
All the methods of assessment tested in our study gave inconsistent results. Levels of intraobserver agreement were moderate and better than levels of interobserver agreement, which were poor or fair. At times interobserver agreement was less than would be expected by chance alone. The expert grade showed no greater degree of agreement than the other grades.
Our results put into question the validity of the radio-
logical grading systems used to assess implant-cement-
bone interfaces. The systems may perform more reliably
when used by those who devised them, but when they are
used at other centres and by different grades of surgeon our
study shows that they are less reliable. The use of trainees
in our study may be criticised on the grounds that they lack
experience in the interpretation of radiographs, but such
grades of surgeon are often responsible for the production
of data in clinical trials.

The results are important and are a cause for concern.
The ‘Johnston’ system forms the basis of a standard system
of terminology for reporting results (Johnston et al 1990).
In the original description of the system some of the
deficiencies were recognised. The authors did not believe
that the true meaning of radiolucency was known and that
the system should be viewed as providing only building
blocks of information. Our results confirm their anxieties
about the definition and interpretation of lucencies.

At first sight, the ‘Barrack’ system, because of its simpli-
city, may appear to be the most reliable, but we found this
not to be so. The grading system has been used to demon-
strate an improvement in the quality of cementing after the
introduction of ‘second-generation’ techniques and has
been used to account for reduced rates of loosening (Bar-
rack et al 1992). In the light of our results this interpretation
should be viewed with some caution as we intended to
compare the two cementation techniques and the study was
designed accordingly. The lack of agreement, however,
meant that such a comparison was unreliable.

The situation must be improved: new methods and sys-
tems should be clearly described in full, with complete
definitions of all terms used, such as lucency. Clinical
training which includes ‘hands-on’ experience is the best
way of sharing terminology and experience. Our assessors
expressed an inability to describe fully and accurately what
they saw. For instance, the term ‘defect’ was commonly
used in preference to the required ‘lucency’. This indicates
that a modified and more appropriate system is required for
assessing the cement mantle. Migration of the prosthesis
and changes in bone density around it may be more useful
as measures for assessment of loosening. Roentgen stereo-
photogrammetric analysis and estimations of bone density
may prove to be more reliable outcome measures in the
future.

We wish to thank the reviewers, Mr C. Heras-Palov, Mr S. Calder, Mr P.
Triffitt, Mr R. Ratnam, Mr J. Timperley and Mr E. Smith.

No benefits in any form have been received or will be received from a
commercial party related directly or indirectly to the subject of this
article.

REFERENCES

Barrack RL, Mulroy RD Jr, Harris WH. Improved cementing tech-
niques and femoral component loosening in young patients with hip

Cohen J. A coefficient of agreement for nominal scales. Educ Psychol

Fleiss JL. Measuring nominal scale agreement among many raters. Psy-
chol Bull 1971;76:378-82.

Gartland JJ. Orthopaedic clinical research: deficiencies in experimental
design and determinations of outcome. J Bone Joint Surg [Am]
1988;70-A:1357-64.

Gruen TA, McNeice GM, Amstutz HC. “Modes of failure” of cemented
stem-type femoral components: a radiographic analysis of loosening.

Johnston RC, Fitzgerald RH, Harris WH, et al. Clinical and radi-
ographic evaluation of total hip replacement: a standard system of

Landis JR, Koch GG. The measurement of observer agreement for

Rook AJ. A Genstat procedure to calculate a kappa coefficient of agree-
ment for nominally scaled data. Genstat Newsletter 1994;31:45-7.

Wright JG, Feinstein AR. Improving the reliability of orthopaedic meas-